Healthy AI First, Folge 8: Ehrlich messen, selbst ausprobieren (Serienfinale)

Quellen und Belege zur Folge

Miriams Bereichsleiter will bis Monatsende eine Zahl: Wie viel schneller ist das Team mit KI? Statt einer großen glänzenden Zahl begleitet die Folge Miriam, Daniel und Jonas vier Wochen lang wie in einem Tagebuch: Woche eins nur beobachten, in Woche zwei und drei arbeitet einer mit der neuen Regel und einer mit der alten, nach einer Woche wird getauscht, Woche vier wird ausgewertet. Gemessen wird mit einem Armaturenbrett aus drei Anzeigen (Minuten pro Änderung, Kraft am Abend, Verständnisgespräch). Unterwegs begegnen uns das Goodhartsche Gesetz, Mitbestimmung und Datenschutz, ein kaputter Donnerstag, den eine KI-Zusammenfassung im Median versteckt, und eine Seite für die Budgetrunde. Zum Schluss ein letztes Mal eure Zahl.

Alle Links wurden am 17.10.2026 geprüft.

Was wir in dieser Folge erwähnen

Die erste Versuchung

- **Aussage:** Wird eine Kennzahl zum Ziel, hört sie auf, eine gute Kennzahl zu sein; der Gedanke geht auf den Ökonomen Charles Goodhart zurück, die bekannte Kurzform stammt von der Anthropologin Marilyn Strathern.
  Quelle: Goodhart, C. (1975): Problems of Monetary Management: The U.K. Experience. Papers in Monetary Economics, Reserve Bank of Australia; Strathern, M. (1997): "Improving ratings": audit in the British University system. European Review 5(3), 305-321.
  Link: https://doi.org/10.1017/S1062798700002660 (Überblick mit Originalwortlaut Goodharts: https://en.wikipedia.org/wiki/Goodhart%27s_law)
  Einordnung: Erfahrungssatz aus Ökonomie und Anthropologie, kein Experiment. Das Beispiel "übernommene Änderungen belohnen" ist eine Anwendung des Prinzips.

Das Armaturenbrett

- **Aussage:** Minuten pro Änderung zählen nur menschliche Zeit (klären, erklären, prüfen, nachbessern); "fertig" heißt: Kriterien erfüllt, Prüfungen bestanden, ein Mensch übernimmt die Verantwortung.
  Quelle: Grundlage der Serie, Kapitel 14 ("menschliche Aufmerksamkeitsminuten je akzeptierter Änderung").
  Einordnung: Ausdrücklich eine Arbeitshypothese, kein etablierter Standard.
- **Aussage:** Mehrere Anzeigen statt einer Zahl: wenige Kennzahlen aus mehreren Dimensionen, bewusst in Spannung, mit mindestens einer Wahrnehmungsmessung, ausgewertet nur auf Teamebene; die Kennzahlen nicht zu Zielvorgaben machen und Teams nicht gegeneinander antreten lassen.
  Quelle: Forsgren, N., Storey, M.-A., Maddila, C., Zimmermann, T., Houck, B. & Butler, J. (2021): The SPACE of Developer Productivity. ACM Queue 19(1), 20-48; DORA: Software delivery performance metrics.

Link: https://doi.org/10.1145/3454122.3454124 (https://queue.acm.org/detail.cfm?id=3454124) · https://dora.dev/guides/dora-metrics/
Einordnung: Forschungsbasierte Messrahmen, keine einzelne Kausalstudie. Im Audio ohne Namen, Details nur in den Shownotes.

Tank und Warnlampe

- **Aussage:** Die Kraft am Abend auf einer Skala von eins bis fünf ist ein lokales Signal, keine Diagnose; NASA-TLX erfasst subjektive Arbeitsbelastung in sechs Dimensionen; das ergänzt die Abendkraft-Skala, misst aber etwas anderes.
  Quelle: Grundlage der Serie, Kapitel 14; NASA Human Systems Integration Division: NASA Task Load Index (TLX); Hart, S. G. (2006): NASA-Task Load Index (NASA-TLX); 20 Years Later. Proceedings of the HFES Annual Meeting 50(9), 904-908.
  Link: https://www.nasa.gov/human-systems-integration-division/nasa-task-load-index-tlx/ · https://doi.org/10.1177/154193120605000909
  Einordnung: Die Skala eins bis fünf ist eine Festlegung der Serie; NASA-TLX ist ein etabliertes Messinstrument für Belastung, keine Ermüdungsdiagnose.
- **Aussage:** Das Verständnisgespräch prüft an einer zufällig gezogenen Änderung, was nie passieren darf und was bei einem Fehler geschieht, den kein Test abdeckt; Selbstvertrauen allein genügt nicht.
  Quelle: Grundlage der Serie, Kapitel 14 und 15.
  Einordnung: Vorschlag, keine validierte Methode. Eine fachlich richtige Erklärung ist ein Hinweis auf Verständnis; das Gespräch ist ein Prüfversuch, kein vollständiger Nachweis.

Woche eins: Nur zuschauen

- **Aussage:** Zeiten und Kraftwerte pro Person sind persönliche Daten; wo es einen Betriebsrat gibt, ist ein technisches System, das Leistung oder Verhalten Einzelner überwachen kann, in Deutschland mitbestimmungspflichtig.
  Quelle: § 87 Abs. 1 Nr. 6 Betriebsverfassungsgesetz; Bundesarbeitsgericht, Beschluss vom 09.09.1975, 1 ABR 20/74 (objektive Eignung zur Überwachung genügt).
  Link: https://www.gesetze-im-internet.de/betrvg/__87.html · https://www.haufe.de/id/entscheidungen/bag-beschluss-vom-09091975-1-abr-2074-HI436774.html
  Einordnung: Gesetzestext und Rechtsprechung; kein Rechtsrat.
- **Aussage:** Claude Code kann auf Wunsch aufzeichnen, wie lange ein Mensch in seiner Oberfläche aktiv ist, und trennt die Zeit, in der jemand selbst eingibt, von der Zeit, in der der Agent arbeitet; ein Zeitstempel ist trotzdem keine Aufmerksamkeit.
  Quelle: Anthropic: Claude Code Dokumentation, Monitoring.
  Link: https://code.claude.com/docs/en/monitoring-usage
  Einordnung: Herstellerdokumentation; misst gebundene Zeit in der Claude-Code-Oberfläche, nicht Aufmerksamkeit oder Belastung. Git liefert Commit-Zeitpunkte, Reviews stehen im Pull-Request-System; zusammen mit dem Kalender ergibt das noch keine vollständigen aktiven Personenminuten. Arbeit außerhalb muss ergänzt, Überschneidungen dürfen nicht doppelt gezählt und verbleibende Messlücken sollten offengelegt werden; Fehler können in beide Richtungen gehen.

Der Seitenwechsel

- **Aussage:** Woche eins Istzustand, Woche zwei und drei Wechsel der Bedingungen (Crossover), Woche vier Auswertung; die Reihenfolge wird verteilt, damit nicht eine leichte oder schwere Woche entscheidet. Ganz sauber ist das nicht, weil Lerneffekte über den Tausch hinweg wirken.
  Quelle: Grundlage der Serie, Kapitel 15.
  Einordnung: Gängiges Versuchsdesign; im kleinen Team ein Lernsignal, kein statistischer Beweis. Das Fußballbild ist unseres.
- **Aussage:** Die neue Regel kommt aus Folge 4: eine komplexe Aufgabe zur Zeit, höchstens ein stiller Hintergrundjob, eine feste Grenze für wartende Prüfungen; Erfolgskriterium und Abbruchgrenzen werden vorher festgelegt.
  Quelle: Grundlage der Serie, Kapitel 7 und 15; DORA: Work in process limits.
  Link: https://dora.dev/capabilities/wip-limits/
  Einordnung: Pilotvorschlag; einen allgemeingültigen Prozentwert für Erfolg gibt es nicht.

Woche drei

- **Aussage:** Wer jede Ersparnis sofort in mehr Arbeit umwandelt, misst etwas anderes; bei gleichem tatsächlichem Durchsatz und stabilen Mittelwerten bedeutet mehr Arbeit im System eine längere mittlere Durchlaufzeit.
  Quelle: Grundlage der Serie, Kapitel 13; Little, J. D. C. (1961): A Proof for the Queuing Formula: L = λW. Operations Research 9(3), 383-387 (ausführlich in Folge 4).
  Link: https://doi.org/10.1287/opre.9.3.383
  Einordnung: Typische Fehlentwicklung laut Grundlage; Little gilt für stabile Systeme im Mittel.

Woche vier: Die glatte Zusammenfassung

- **Aussage:** Ein eintägiger Ausreißer in einigen Dutzend Werten taucht im Median kaum auf. Der Mittelwert steigt deutlich, weil er den echten Reparaturaufwand mitrechnet; das ist ohne Messfehler keine Verzerrung. Deshalb zusätzlich die Verteilung und den gesamten menschlichen Aufwand zeigen. Eine KI-Zusammenfassung ist keine Prüfung.
  Quelle: Statistisches Grundwissen; Grundlage der Serie, Kapitel 13 bis 15 ("Verteilungen und Ausreißer statt nur einen Durchschnitt").
  Einordnung: Allgemeine Statistik; der kaputte Donnerstag ist erfunden.

Eine Seite für die Budgetrunde

- **Aussage:** Was gemessen wurde (dokumentierte Zeiten, ergänzte Tätigkeiten und bekannte Messlücken; Kraft und Verständnis von den Menschen), was sich zeigte, was offen bleibt und welche nächste Entscheidung ansteht.
  Quelle: Grundlage der Serie, Kapitel 15 und 16 (Vorlagen).
  Einordnung: Vorlage; das Ergebnis von Miriams Versuch ist erfunden.

Was noch niemand weiß

- **Aussage:** Dass Wachsamkeit anstrengt, ist lange bekannt; wie gut ein Review am späten Nachmittag nach einem Tag voller Agentenarbeit noch ist, dafür ist uns keine belastbare direkte Messung bekannt. Offen ist auch, ob KI Ressourcen freisetzt oder die erwartete Arbeit einfach mitwächst.
  Quelle: Warm, J. S., Parasuraman, R. & Matthews, G. (2008): Vigilance Requires Hard Mental

Work and Is Stressful. Human Factors 50(3), 433-441; Grundlage der Serie, Kapitel 17.
Link: https://doi.org/10.1518/001872008X312152
Einordnung: Vigilanzforschung als Grundlage; die Übertragung auf Code Review ist eine offene Forschungsfrage.

Eure Zahl

- **Aussage:** In Folge 1 stand Daniel für sechzehn erfahrene Entwickler, die sich fast ein Viertel schneller schätzten und gemessen langsamer waren.
  Quelle: Becker, J., Rush, N., Barnes, E. & Rein, D. (METR, 2025): Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity. Preprint.
  Link: https://arxiv.org/abs/2507.09089 (Zusammenfassung: https://metr.org/blog/2025-07-10-early-2025-ai-experienced-os-dev-study/)
  Einordnung: Randomisierte Studie, 16 Personen, Werkzeuge von Anfang 2025.

Werkzeug-Tipp: Zeiten automatisch erheben

- Claude Code Monitoring über OpenTelemetry: Aktivierung mit `CLAUDE_CODE_ENABLE_TELEMETRY=1` plus Exporter-Konfiguration (standardmäßig aus). Die Metrik `claude_code.active_time.total` misst aktive Zeit in Sekunden, Attribut `type` = `user` (Tastatureingaben) oder `cli` (Werkzeugausführung und KI-Antworten). https://code.claude.com/docs/en/monitoring-usage
- Zeiten aus Pull Requests und Commits (siehe Folge 2): Git liefert Commit-Zeitpunkte, Reviews und Merges stehen im Pull-Request-System. Das sind Durchlaufzeiten, keine aktiven Personenminuten. https://cli.github.com/manual/gh_pr_list · https://git-scm.com/docs/git-log

Zum Weiterlesen

- DORA: A history of DORA's software delivery metrics. Deployment Rework Rate kam 2024 als fünfte Kennzahl hinzu (ungeplante Auslieferungen infolge eines Produktionsvorfalls, nicht identisch mit interner Nacharbeit). https://dora.dev/insights/dora-metrics-history/
- Noda, A., Storey, M.-A., Forsgren, N. & Greiler, M. (2023): DevEx: What Actually Drives Productivity. ACM Queue 21(2), 35-53. https://doi.org/10.1145/3595878
- Levitt, S. D. & List, J. A. (2011): Was There Really a Hawthorne Effect at the Hawthorne Plant? An Analysis of the Original Illumination Experiments. American Economic Journal: Applied Economics 3(1), 224-238. Die berühmte Fabrikgeschichte ist wackliger als ihr Ruf. https://doi.org/10.1257/app.3.1.224 (Working Paper: https://www.nber.org/papers/w15016)
- Paradis, E. et al. (2024): How much does AI impact development speed? An enterprise-based randomized controlled trial. 96 Google-Softwareentwickler; nach bester Schätzung etwa 21 % weniger Zeit, großes Konfidenzintervall. Preprint. https://arxiv.org/abs/2410.12944
- Grundlage der Serie: Healthy AI First Software Development, Fassung 2 (01.10.2026), Kapitel 12 bis 17.

Hinweis

Healthy AI First ist KI-recherchiert und KI-generiert, gesprochen von einer KI-Version der Stimme von Björn und von ihm geprüft.

Miriam, Daniel, Jonas und der Bereichsleiter sind fiktive Figuren zur Veranschaulichung. In Folge 1 stand Daniel stellvertretend für die Teilnehmenden der METR-Studie; ab Folge 2 ist er ein erfundener Kollege und hat an der Studie nicht teilgenommen. Auch Verlauf und Ergebnisse von Miriams Versuch (die schwere Woche zwei, der kaputte Donnerstag, das Verständnisgespräch) sind erfunden.